Of Stacks and Muses: Adventures in Learning Analytics at Marist College

Eitel Lauría, Edward Presutti, Maria Kapogiannis

To cite this version:

Eitel Lauría, Edward Presutti, Maria Kapogiannis. Of Stacks and Muses: Adventures in Learning Analytics at Marist College. LatinX in AI Research at ICML 2019”, Jun 2019, Long Beach, CA, United States. hal-02265832

HAL Id: hal-02265832
https://hal.archives-ouvertes.fr/hal-02265832
Submitted on 12 Aug 2019

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
MUSE Databases

MUSE provides predictions of academic success of undergraduate students in a given course, six weeks into the semester of a 15-week course. The threshold of good academic standing is a letter grade C (students with less than a C are considered at risk). To make predictions less tiresome, we tie this prediction to a data-driven binary variable (the aforementioned threshold, and we use a color coding (see Figure 2).

- **GREEN** for students in good standing (those with a probability of success $\geq 55\%$)
- **YELLOW** for student with an undetermined risk status (probability of success between 55% and 45%)
- **RED** for at-risk students (those with a probability of success $< 45\%$)

Figure 2. Database with Predictions

### Stacked ensemble of classifiers

Training and testing a two-stage stack with 8 binary classifiers in the first stage, one binary classifier in the second stage, and 3 independent data sets A, B, C, to avoid data leakage (see Figure 3). After the stack is trained, tuned, and tested, it can be used to make predictions on new data D. Figure 4 depicts the two-stage stack making predictions on incoming (and therefore unknown) data D.

Figure 3. Training and testing a two-stage stack

After the stack is trained, tuned, and tested, it can be used to make predictions on new data D. Learning algorithms used to train base models should be as diverse as possible so that the predictions made by them have relatively low correlations ($\leq 0.75 \leftrightarrow 0.80$). If predictions are highly correlated, it indicates that the base models map very similar hypothesis functions, which defeats the purpose of using a stack.

Figure 4. Using the stack for prediction on new data

We studied the use of a two-stage stack which draws MUSE-trained with undergraduate data from 10 semesters (Fall 2012 – Spring 2017). LMS student activity data is recorded as weekly frequency ratios, normalized with the mean and std. dev. of each course. The target variable is the final grade, recorded as a binary variable - Academic Risk - using letter grade C as threshold (see Table 2). A random sample of 51029 records (35% of the total) was used. We performed 8 batches of experiments, using 2 different configurations of classification algorithms for the first stage (base models), 2 different sets of predictors to train the base models; and 2 different algorithms for the second-stage model. Each batch was repeated twice in 3 runs with 10-fold random generator seeds to account for variation in predictive performance due to the date; in each run the data was randomly partitioned into datasets A, B and C with 10345 records each, and augmented to a total of 80 runs in the experiment ($2 \times 2 \times 3 \times 10$). See Table 3-5 for details.

### Experimental Setup

We used the study of a two-stage stack which draws MUSE-trained with undergraduate data from 10 semesters (Fall 2012 – Spring 2017). LMS student activity data is recorded as weekly frequency ratios, normalized with the mean and std. dev. of each course. The target variable is the final grade, recorded as a binary variable - Academic Risk - using letter grade C as threshold (see Table 2). A random sample of 51029 records (35% of the total) was used. We performed 8 batches of experiments, using 2 different configurations of classification algorithms for the first stage (base models), 2 different sets of predictors to train the base models; and 2 different algorithms for the second-stage model. Each batch was repeated twice in 3 runs with 10-fold random generator seeds to account for variation in predictive performance due to the date; in each run the data was randomly partitioned into datasets A, B and C with 10345 records each, and augmented to a total of 80 runs in the experiment ($2 \times 2 \times 3 \times 10$). See Table 3-5 for details.